Journal of Toxicology and Environmental Health, Part A
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713667303

Evaluating Sufficient Similarity for Drinking-Water Disinfection By-Product (DBP) Mixtures with Bootstrap Hypothesis Test Procedures

Paul I. Feder [a]; Zhenxu J. Ma [a]; Richard J. Bull [b]; Linda K. Teuschler [c]; Glenn Rice [c]
[a] Battelle, Statistics and Information Analysis, Columbus, Ohio, USA [b] MoBull Consulting, Richland, Washington, USA [c] U.S. Environmental Protection Agency, Cincinnati, Ohio, USA

Online Publication Date: 01 January 2009

To cite this Article Feder, Paul I., Ma, Zhenxu J., Bull, Richard J., Teuschler, Linda K. and Rice, Glenn(2009)'Evaluating Sufficient Similarity for Drinking-Water Disinfection By-Product (DBP) Mixtures with Bootstrap Hypothesis Test Procedures',Journal of Toxicology and Environmental Health, Part A,72:7,494 — 504

To link to this Article: DOI: 10.1080/15287390802608981

URL: http://dx.doi.org/10.1080/15287390802608981

Journal of Toxicology and Environmental Health, Part A, 72: 494–504, 2009

ISSN: 1528-7394 print / 1087-2620 online
DOI: 10.1080/15287390802608981

Taylor & Francis
Taylor & Francis Group

Evaluating Sufficient Similarity for Drinking-Water Disinfection By-Product (DBP) Mixtures with Bootstrap Hypothesis Test Procedures

Paul I. Feder[1], Zhenxu J. Ma[1], Richard J. Bull[2], Linda K. Teuschler[3], and Glenn Rice[3]

[1]*Battelle, Statistics and Information Analysis, Columbus, Ohio,* [2]*MoBull Consulting, Richland, Washington, and* [3]*U.S. Environmental Protection Agency, Cincinnati, Ohio, USA*

In chemical mixtures risk assessment, the use of dose-response data developed for one mixture to estimate risk posed by a second mixture depends on whether the two mixtures are sufficiently similar. While evaluations of similarity may be made using qualitative judgments, this article uses nonparametric statistical methods based on the "bootstrap" resampling technique to address the question of similarity among mixtures of chemical disinfectant by-products (DBP) in drinking water. The bootstrap resampling technique is a general-purpose, computer-intensive approach to statistical inference that substitutes empirical sampling for theoretically based parametric mathematical modeling. Nonparametric, bootstrap-based inference involves fewer assumptions than parametric normal theory based inference. The bootstrap procedure is appropriate, at least in an asymptotic sense, whether or not the parametric, distributional assumptions hold, even approximately. The statistical analysis procedures in this article are initially illustrated with data from 5 water treatment plants (Schenck et al., 2009), and then extended using data developed from a study of 35 drinking-water utilities (U.S. EPA/AMWA, 1989), which permits inclusion of a greater number of water constituents and increased structure in the statistical models.

Many contaminants occur in the environment as complex mixtures, which are generally comprised of many, perhaps hundreds, of chemical components. The proportions of these components vary depending on how the mixture was formed and the fate of the mixture in the environment. Furthermore, a portion of the complex mixture may be poorly characterized chemically, with some of its components not known.

The U.S. Environmental Protection Agency (U.S. EPA) issued guidance documents (*Guidelines for the Health Risk Assessment of Complex Mixtures*, hereafter the *Guidelines* [U.S. EPA, 1986], *Supplementary Guidance for Conducting Health Risk Assessment of Complex Mixtures* [U.S. EPA, 2000]) that discuss concepts and suggested procedures for the assessment of health risks associated with exposures to multiple chemicals, including complex mixtures. The U.S. EPA (2000) defines a complex mixture as a mixture containing many, perhaps hundreds of, components. The chemical composition may vary over time or with different conditions under which the mixture is produced. Complex mixture components may be generated simultaneously as by-products from a single source or process, such as the disinfection of drinking water at water treatment plants, intentionally produced as a commercial product, or may coexist because their chemical properties result in their co-occurrence in an environmental medium.

The *Guidelines* (U.S. EPA, 1986) state that the preferred situation for assessing health risks associated with exposure to complex mixtures is when compositional and toxicity data are directly available on the mixture of concern. Procedures can then be adopted that are similar to those used to evaluate risks associated with single compounds. However this preferred situation does not always occur. To deal with such situations, the *Guidelines* state:

> If no data are available on the mixture of concern, but health effects data are available on a similar mixture . . . a decision must be made whether the mixture on which health effects are available is "sufficiently" similar to the mixture of concern to permit a risk assessment . . . . In determining reasonable similarity, consideration should be given to any information on the components that differ . . . between the mixture on which health effects are available and the mixture of concern.

The views expressed in this article are those of the individual authors and do not necessarily reflect the views and policies of the U.S. Environmental Protection Agency (EPA). Those sections prepared by U.S. EPA scientists have been reviewed in accordance with U.S. EPA peer and administrative review policies and approved for presentation and publication. Mention of trade names or commercial products does not constitute endorsement or recommendations for use.

The authors acknowledge and appreciate the many helpful review comments and suggestions of Richard C. Hertzberg, Ph. (Emory University), and David Farrar, PhD (U.S. EPA/ORD/NCEA). These comments greatly improved this article.

Address correspondence to Paul I. Feder, Battelle, Statistics and Information Analysis, 505 King Avenue, Columbus, OH 43201-2693, USA. E-mail: feder@battelle.org

Bull et al. (2009a, 2009b) discuss in detail the various classes of disinfection by-products (DBP) that occur in drinking water and are associated with various adverse health consequences. Bull et al. (2009a, 2009b) characterize input factors to the water treatment process that affect the formation of DBP, the variation in the output DBP associated with these factors, and some of the toxicological implications of this variation.

Rice et al. (2009) present an overview of proposed methods for assessing similarity of complex mixtures. Rice et al. (2008) discuss methods based on whole mixtures and methods based on individual mixture component data. They state that if whole-mixtures data are available then whole-mixtures methods are preferred because they are subject to less uncertainty. The statistical methods discussed in this article are based on whole-mixtures data.

Feder et al. (2009) discuss statistical considerations in the comparison of similar mixtures. Multivariate statistical procedures are applied to determine whether individual water supplies exhibit statistically significant variation from a group of water supplies that are considered to be similar. Principal components analysis is applied as a dimensionality reduction and data visualization technique. The statistical procedures are illustrated with examples based on the Schenck et al. (2009) data.

This article extends Feder et al. (2009) in several directions:

1. The statistical methods (Feder et al., 2009) are based on normal theory multivariate analysis of variance. This article presents nonparametric alternatives to the normal theory analysis procedures based on the nonparametric "bootstrap" technique (Westfall & Young, 1993; Efron & Tibshirani, 1993). The bootstrap resampling technique is a general-purpose, computer-intensive approach to statistical inference that substitutes empirical sampling for theoretically based parametric mathematical modeling. Nonparametric, bootstrap-based inference involves fewer assumptions than parametric normal theory-based inference. It produces valid results, at least in an asymptotic sense, whether or not the parametric distributional assumptions hold, even approximately. The nonparametric procedures are less dependent on distributional assumptions and apply to a wide variety of inference situations. The bootstrap counterpart to normal theory multivariate analysis of variance is discussed in detail and is illustrated by example.
2. The statistical analysis procedures are illustrated with the Schenck et al. (2009) study data as well as with data based on the 35-utility study, conducted by the U.S. Environmental Protection Agency and the Association of Metropolitan Water Districts (U.S. EPA/AMWA, 1989). The multivariate statistical analysis procedures in Feder et al. (2009) are extended to the 35-utility data set. Multivariate analysis of variance and its bootstrap counterparts are applied to assess similarity of treated water DBP chemical components among alternative water sources, water treatments, seasons, and utilities.

The remainder of the article discusses and illustrates the bootstrap technique and its application to assessments of similarity among water sources.

It should be noted, as discussed in Feder et al. (2009), that although most of the water characteristics used to compare treatments and water samples in the examples in this article are chemical characteristics, the similarity of mixtures with respect to toxicological effects is of principal interest from a public health perspective. Thus, the similarity of chemical characteristics among mixtures is of interest primarily to the extent that the similarity of chemical characteristics implies the similarity of toxicological characteristics.

APPLICATION OF BOOTSTRAP ANALYSIS TO THE ASSESSMENT OF SIMILARITY OF DBPs IN ALTERNATIVE WATER SUPPLIES

The bootstrap sampling method (Efron & Tibshirani, 1993; Westfall & Young, 1993) is a general, flexible, computationally intensive approach to inference that is relatively insensitive to sampling assumptions concerning underlying distributions. Comparisons among classes of water supplies based on normal theory multivariate analysis of variance (Feder et al., 2009) are reconsidered based on the bootstrap technique.

The inference procedures in Feder et al. (2009) are illustrated with the Schenck et al. (2009) study data. These data originate from five treatment plants, each of which was monitored at a single time point between 1995 and 1998. The inference procedures are illustrated based on a seven-dimensional vector of output water chemical composition:

- Total organic carbon (TOC) (mg/L).
- Total organic halogens (TOX) (μg/L).
- Mutagenic activity (revertants/L equivalent).
- Total trihalomethanes (TTHM) (μg/L).
- HAA6 (six haloacetic acids[1]) (μg/L).
- Percent brominated TTHM.
- Percent brominated HAA6.

These "summary metrics" are important to characterizing the complex mixture as a whole, as opposed to only evaluating a defined set of chemical components (e.g., only looking for similar levels of individual components in the mixture such as chloroform or dichloroacetic acid). These summary metrics account for unknown chemicals in the complex mixture, compositional changes, and possible interactions among chemicals. These seven output water characteristics were recorded at each treatment plant for the finished water at the plant and for the distribution system water at two locations away from the plant. Two-sample and matched-pairs analysis examples are illustrated with this data set.

[1]The six haloacetic acids denoted as HAA6 are chloroacetic acid, bromoacetic acid, dichloroacetic acid, trichloroacetic acid, bromochloroacetic acid, and dibromoacetic acid.

TWO-SAMPLE ANALYSIS

Normal Theory Test

For an initial illustration with a relatively simple example, the five treatment plants are treated as if they are replicates of a single process (i.e., plant identification is ignored). The purpose of the analysis is to determine whether the distribution water samples are sufficiently similar to the finished water samples. Let N_X denote the number of finished water samples, N_Y denote the number of distribution water samples, and p denote the dimensionality of the response vectors. In this example, $N_X = 5$, $N_Y = 10$, and $p = 7$. Denote the response vectors for the finished water by $X_1, X_2, \ldots, X_5$ and the response vectors for the distribution water by $Y_1, Y_2, \ldots, Y_{10}$. It is assumed that the X_i terms and the Y_i terms are independent with mean and covariance $X \sim (\mu_X, \Sigma_X)$; $Y \sim (\mu_Y, \Sigma_Y)$. Since there are just $N_X = 5$ finished water samples, there are not sufficient replicates to estimate separate covariances within categories with a seven-dimensional response vector or to evaluate the assumption of variance homogeneity. It is assumed that $\Sigma_X = \Sigma_Y \equiv \Sigma$.

The sample means and the pooled sample covariance matrix are estimated as

$$\bar{X} = \frac{\sum_{i=1}^{N_X} X_i}{N_X}; \qquad \bar{Y} = \frac{\sum_{i=1}^{N_Y} Y_i}{N_Y}$$

$$S = \frac{\sum_{i=1}^{N_X} (X_i - \bar{X})(X_i - \bar{X})' + \sum_{i=1}^{N_Y} (Y_i - \bar{Y})(Y_i - \bar{Y})'}{(N_X - 1) + (N_Y - 1)}$$

where S has $\nu \equiv [(N_X - 1) + (N_Y - 1)]$ degrees of freedom.

The hypothesis to be tested is

$$H_0: \mu_X = \mu_Y$$

versus

$$H_1: \mu_X \neq \mu_Y$$

The normal theory test procedure is based on Hotelling's T^2 statistic (Morrison, 1976, p. 128ff).[2]

$$T_o^2 = (\bar{X} - \bar{Y})' \left[\left(\frac{1}{N_X} + \frac{1}{N_Y} \right) S \right]^{-1} (\bar{X} - \bar{Y})$$

[2]The reference citations refer to Morrison (1976), *Multivariate Statistical Methods*, second edition. There is now a fourth edition of this book available (2002); however, the second-edition citations were retained to maintain the page references.

Under the null hypothesis, the distribution of T_0^2 is proportional to an F distribution, namely,

$$\frac{v - (p-1)}{pv} T_0^2 \sim F_{p, v-(p-1)}$$

Nonparametric Bootstrap Test

To carry out the nonparametric bootstrap resampling counterpart to the normal theory procedure, a random sample of multivariate responses that reflects H_0 is drawn from an approximation to the underlying distribution under the null hypothesis: $F_X = F_Y \equiv F$ (where F_X and F_Y denote the cumulative distribution functions of the vectors X and Y, respectively). An empirical reference distribution is developed for the test statistic based on the random sample. The statistical significance of the observed statistic for the actual sample data is evaluated based on this reference distribution. Hall and Wilson (1991) present guidelines for characteristics that should be possessed by bootstrap resamples:

- Even if the data may be drawn from a population that fails to satisfy H_0, the resampling should be done in a way that reflects H_0.
- Extraneous parameters (usually scale parameters) should be eliminated from the bootstrap hypothesis test statistic to the extent possible.

To carry out the resampling process in the two-sample case for the Schenck et al. (2008) data, it is necessary to construct a sample that reflects H_0, at least in an asymptotic sense:

- Fit a two-sample model to the data (e.g., with a one-way multivariate analysis of variance with two groups) and calculate standardized residuals $\{\varepsilon_{Xi}\}$, $\{\varepsilon_{Yj}\}$, where

$$\varepsilon_{Xi} = (X_i - \bar{X}) \sqrt{\frac{N_X}{N_X - 1}}$$

$$\varepsilon_{Yi} = (Y_i - \bar{Y}) \sqrt{\frac{N_Y}{N_Y - 1}}$$

The $\{\varepsilon_{Xi}\}$, $\{\varepsilon_{Yi}\}$ are approximately independent $(0, \Sigma)$ as N_X, $N_Y \to \infty$. We then form a sample of residuals of size $N_X + N_Y$. The empirical distribution based on the pooled sample is taken as an estimate of the population distribution, assuming the null hypothesis is true.

- Draw a random sample of size N_X, with replacement from the $N_X + N_Y$ ε's, and call them $\varepsilon^*_{X1}, \ldots, \varepsilon^*_{XNx}$.
- Draw a random sample of size N_Y, with replacement from the $N_X + N_Y$ ε's, and call them $\varepsilon^*_{Y1}, \ldots, \varepsilon^*_{YNy}$.
- Calculate $\bar{\varepsilon}^*_X, S_X^*, \bar{\varepsilon}^*_Y, S_Y^*$ and use these values to calculate the counterpart to the Hotelling T^2 statistic:

$$T^{*2} = (\bar{\varepsilon}^*_X - \bar{\varepsilon}^*_Y)' [(\tfrac{1}{N_X} + \tfrac{1}{N_Y}) S^*_\varepsilon)]^{-1} (\bar{\varepsilon}^*_X - \bar{\varepsilon}^*_Y)$$

where S_E^* is a pooled estimate of variability between the two groups. T^{*2} has the null distribution of Hotelling's T^2 statistic (asymptotically as N_X, $N_Y \to \infty$) irrespective of whether H_0 is true.

- Generate B (e.g., $B = 1000$) resamples $\{\varepsilon_X^*\}$, $\{\varepsilon_Y^*\}$ from the residuals with replacement, and for each resample, calculate T^{*2}.
- Form the bootstrap reference distribution of T^{*2} from the realizations of the B resamples.
- Compare the observed value of T_0^2 (used in the normal theory test) to this reference distribution.
- The bootstrap significance level equals the number of T^{*2} values that exceed T_0^2, divided by B.

Note that this approach resamples the residuals $\{\varepsilon_X\}$, $\{\varepsilon_Y\}$ rather than the X's and Y's directly, because the resampling distribution of the X's and Y's would not reflect the null distribution if, in fact, the null hypothesis did not hold. In such a case, both the numerator $\overline{X}^* - \overline{Y}^*$ and the denominator S^* would be inflated.

Example—Schenck et al. (2009) Study

For the two-sample analysis based on the Schenck et al. (2008) data, $p = 7$ and $\nu - (p - 1) = (15 - 2) - (7 - 1) = 7$. Thus, the normal theory Hotelling's T^2 test is based on an F distribution with 7 numerator and 7 denominator degrees of freedom. The Hotelling's T^2 test is carried out by performing multivariate analysis of variance using the GLM procedure within the SAS System (SAS Institute, Inc., 2004). The results of the SAS analysis are displayed in Table 1.

The multivariate analysis of variance option within the GLM procedure reports the results of four test criteria based on the nonzero characteristic roots of the hypothesis and error variance covariance matrix "ratio," HE^{-1}, which generalize the univariate analysis of variance F statistic. The "Hotelling–Lawley Trace" is the sum of the characteristic roots. "Roy's Largest Root" is the largest of the characteristic roots. "Pillai's Trace" is ["Hotelling–Lawley Trace"]/[1 + "Hotelling–Lawley Trace"].[3] "Wilks's Lambda" is the reciprocal of the product of the characteristic roots of $HE^{-1} + I$ (i.e., 1 + characteristic roots of HE^{-1}).

The number of nonzero characteristic roots is at most the number of linearly independent contrasts, s, in the hypothesis space. In the case of the two-sample comparison, $s = 1$. In this special case, only one positive characteristic root of HE^{-1} exists. Let λ (= 0.954 in Table 1) denote the single positive characteristic root of HE^{-1}. The Hotelling–Lawley Trace and Roy's Greatest Root are both equal to λ. Pillai's Trace is $\lambda/(1 + \lambda)$. Wilks's Lambda is $1/(1 + \lambda)$. The four criteria are equivalent to one another, as reflected in the common F values and common significance levels in Table 1. Furthermore, in this case, they are multiples of and equivalent to the two-sample Hotelling's T^2 statistic.

Table 1 shows that the value of the F statistic associated with the observed value of the Hotelling–Lawley trace is 0.95, which is significant at the .52 level when the normal theory test is applied. (As just discussed, all four statistics in Table 1 are equivalent.)

The results of 999 bootstrap resamples are summarized in Table 2. For each bootstrap resample under the null hypothesis distribution, the multivariate analysis of variance was run and the F value associated with the Hotelling–Lawley Trace was determined. A portion of the listing of the empirical cumulative distribution function is shown in the F value column in Table 2. The location of the observed F value, 0.95, is shown in bold. The observed bootstrap significance level, determined from the Cumulative Frequency column of the table, is 1 – (508/999) = 1 – .51 = .49. This is in close agreement with the normal theory significance level, .52, shown in Table 1.

The empirical distribution function of the F values corresponding to the bootstrap resamples is displayed in Figure 1, along with the normal theory F distribution that is included for comparison. The empirical bootstrap distribution is

[3]This relation between Pillai's Trace and the Hotelling–Lawley Trace is applicable in the special case when there is a single positive characteristic root.

TABLE 1

Hotelling's T^2 Test Statistic for Equality of Finished Water and Distribution Water. One-way normal theory multivariate analysis of variance with two groups. Schenck and Sivaganesan (2008) data.

	F-Statistics for the hypothesis of no overall sample effect				
Statistic	Value	F-Value	Numerator degree of freedom	Denominator degree of freedom	Significance level
Wilks' Lambda	0.512	0.95	7	7	0.52
Pillai's Trace	0.488	0.95	7	7	0.52
Hotelling-Lawley Trace	0.954	0.95	7	7	0.52
Roy's Greatest Root	0.954	0.95	7	7	0.52

TABLE 2
Summary of F-Values Corresponding to $B = 999$ Bootstrap Resamples of the Hotelling's T^2 Test Statistic for Equality of Finished Water and Distribution Water. One-way nonparametric multivariate analysis of variance. Schenck and Sivaganesan (2009) data.

F-Value[a,b]	Cumulative Frequency	Cumulative %
0.86	440	44.04
0.87	445	44.54
0.88	452	45.25
0.89	461	46.15
0.90	476	47.65
0.91	479	47.95
0.92	485	48.55
0.93	495	49.55
0.94	504	50.45
0.95	**508**	**50.85**
0.96	511	51.15
0.97	516	51.65
0.98	518	51.85
0.99	527	52.75
1.00	554	55.46
1.01	558	55.86
1.02	559	55.96

[a]Empirical distribution of ordered Hotelling-Lawley trace F-values based on bootstrap resampling iterations under the null hypothesis.

[b]F-Value of Hotelling-Lawley value based on original data (from Table 1) = 0.95.

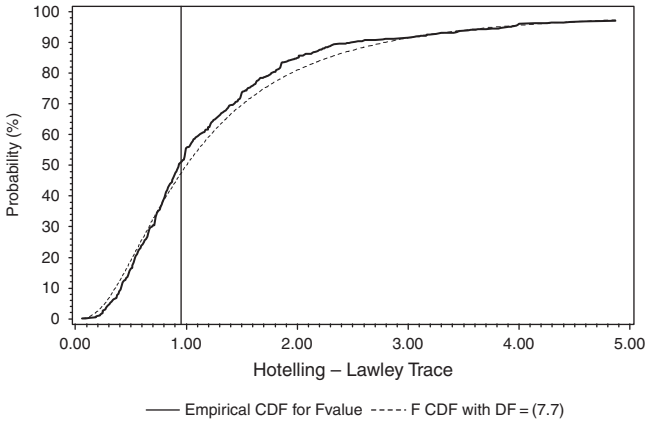

FIG. 1. Empirical CDF and normal theory F-distribution CDF for Hotelling-Lawley trace F-value for testing no sample effects. Two-sample analysis. Significance levels (%) are 100 minus the values at which the vertical reference line crosses the curves. Schenck and Sivaganasen (2009) data.

similar to the normal theory F distribution. A reference line is shown at the F value associated with the observed value of the statistic, .95 (Table 1). The attained significance levels (%) are obtained from the vertical axis as 100 minus the values at which the reference line intersects the cumulative distribution functions.

MATCHED-PAIRS ANALYSIS

Normal Theory Test

A matched-pairs analysis accounts for treatment-plant effects as well as an effect for finished water versus distribution water. Water samples obtained from the same plant are assumed to be related to one another through a common plant effect. Recall that the Schenck et al. (2009) data include a single finished water sample and two distribution water samples for each of five treatment plants. The assumption is made in this analysis that the two distribution system samples from the same plant are replicates, even though they were collected at different locations. A two-way analysis of variance model is used that includes "treatment" effects (i.e., finished water versus distribution water) and "block" effects (i.e., treatment plants).

Let X_{ijk} denote the input and output water characteristics of the water sample. The indices i, j, k have ranges $i = 1,\ldots,I$ for "treatment" effects ($I = 2$), $j = 1,\ldots,J$ for "block" (treatment plant) effects ($J = 5$), and $k = 1,\ldots,N_{ij}$ for replicates ($N_{1j} = 1$, $N_{2j} = 2$) for all j. It is assumed that

$$X_{ijk} \sim (\mu_{ij}, \Sigma)\; i=1,\ldots,I;\; j=1,\ldots,J;\; k=1,\ldots,N_{ij}$$

A common covariance matrix is assumed across the treatments and treatment plants. As discussed earlier for the two-sample case, the data do not include a sufficient number of replicate determinations to estimate separate covariances or to detect departures from the constant covariance assumption.

As discussed in Feder et al. (2009), the five treatment plants are not considered to be a representative sample from a population of treatment plants. Inferences are to be restricted to the five treatment plants in the data.

The most general two-way analysis of variance model is

$$\mu_{ij} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad i=1,\ldots,I;\; j=1,\ldots,J$$

subject to the conditions $\Sigma_i\alpha_i = \Sigma_j\beta_j = \Sigma_i(\alpha\beta)_{i\,j} = \Sigma_j(\alpha\beta)_{ij} = 0$. The $\{\alpha_i\}$ represent the "treatment" main effects, the $\{\beta_j\}$ represent the "block" main effects, and the $\{(\alpha\beta)_{ij}\}$ represent the "treatment" by "block" interactions. The constraints imply that μ represents the average response across water types and treatment plants, $\mu_\eta + \alpha_i$ represents the average response for water type i averaged across treatment plants, and $\mu_\eta + \beta_j$ represents the average response for treatment plant j, averaged across water types.

The Schenck et al. (2009) data are not sufficient to estimate and make inferences about the full model and still be able to estimate the residual covariance. Therefore, an additive main

effects submodel was fitted to the Schenck et al. (2009) data by setting the interactions $\{(\alpha\beta)_{ij}\}$ to zero. The main effects model is:

$$\mu_{ij} = \mu_0 + \alpha_i + \beta_j \quad i = 1,\ldots,I;\ j = 1,\ldots,J$$

with $\Sigma_i \alpha_i = \Sigma_j \beta_j = 0$. This is a matched-pairs model. The treatment plants are the "blocks" and the finished water versus distribution water differences are the "treatment effects," which are specified by the model to be consistent across treatment plants.

The hypotheses

$$H_0^{(\alpha)}: \alpha_1 = \alpha_2$$

versus

$$H_1^{(\alpha)}: \alpha_1 \neq \alpha_2$$

and

$$H_0^{(\beta)}: \beta_j = 0 \text{ for all } j$$

versus

$$H_1^{(\beta)}: \beta_j \neq 0 \text{ for some j}$$

can be tested based on this model. $H_0^{(\alpha)}$ is the hypothesis of no "treatment" effects, while $H_0^{(\beta)}$ is the hypothesis of no "block" (treatment plant) effects. These hypothesis tests can be carried out based on normal theory multivariate analysis of variance tests (Morrison, 1976, p. 177ff).

Nonparametric Bootstrap Test

In analogy with the discussion of the bootstrap procedure for the two-sample analysis, the matched-pairs bootstrap analysis involves drawing repeated random samples from the empirical distribution of the data under the null hypothesis based on the residuals from the two-way multivariate analysis fit and generating null empirical distributions for the test statistics.

To carry out the resampling process in the matched pairs case, it is necessary to construct a sample that reflects H_0, at least in an asymptotic sense:

- Fit the two-way multivariate analysis of variance model and calculate the studentized residuals from the fit:

$$\varepsilon_{ijk} = \left[diag(stderr(X_{ijk} - \hat{X}_{ijk})) \right]^{-1} (X_{ijk} - \hat{X}_{ijk})$$

where $\hat{X}_{ijk}$ denotes the model-predicted value and $diag(stderr(X_{ijk} - \hat{X}_{ijk}))$ is a diagonal matrix with diagonal elements the standard errors of the components of the residual vector.

- The residual vectors are approximately distributed as $\varepsilon_{ijk} \sim (0, P)$ where P is the correlation matrix of the X_{ijk}'s, as $N_X, N_Y \to \infty$. This distribution is approximate in small samples.
- For the Schenck et al. (2008) study, there are $N_X + N_Y \equiv N = 15$ studentized residuals. Draw a random sample of size $N = 15$ with replacement from the studentized residuals. (There are 15^{15} such possible random samples.) Denote these residuals by $\varepsilon 1^*, \varepsilon 2^*, \ldots, \varepsilon 15^*$. These residuals satisfy the null hypothesis.
- With each residual, associate each combination of treatment plant and treatment as follows:
  - $\varepsilon 1^*$ corresponds to Treatment Plant 1, finished water
  - $\varepsilon 2^*$, $\varepsilon 3^*$ correspond to Treatment Plant 1, distribution water
  - $\varepsilon 13^*$ corresponds to Treatment Plant 5, finished water
  - $\varepsilon 14^*$, $\varepsilon 15^*$ correspond to Treatment Plant 5, distribution water
- Calculate the F statistics from the multivariate analysis of variance, F^*, based on these residuals to test the hypotheses of homogeneity among treatments and homogeneity among treatment plants.
- Repeat the resampling process B times (e.g., $B = 1000$) and generate a bootstrap empirical distribution function for F^* that is associated with the multivariate analysis of variance test statistic (e.g., the Hotelling–Lawley Trace statistic).
- The bootstrap significance level is determined as the number of F^* values that exceed F_0, divided by B, where F_0 is the F value associated with the statistic in the original sample.

Example—Schenck et al. (2009) Study

Test for Equality of Finished Water and Distribution Water

For the two-way matched-pairs analysis of no treatment effects based on the Schenck et al. (2009) data, $p = 7$ and $\nu - (p - 1) = [15 - (1 + 1 + 4) - (7 - 1)] = 3$. Thus, the normal theory multivariate analysis of variance test of equality of treatment effects is based on an F distribution with 7 numerator and 3 denominator degrees of freedom. The test is carried out by performing multivariate analysis of variance using the GLM procedure of the SAS System.

The F value associated with the observed value of the Hotelling-Lawley trace is 1.01 (not shown), which is significant at the .55 level. This result agrees with the two-sample analysis.

The bootstrap resample analysis of the multivariate analysis of variance test statistic for equality of finished water and distributed water results in an observed bootstrap significance level approximately 1 – 0.53 = 0.47. This agrees with the normal theory significance level, .55, and is in good agreement with the two-sample analysis.

Test for Equality of Treatment Plants

For tests of equality among the five treatment plants, the hypothesis space has four linearly independent contrasts. The error sums of squares and cross products matrix E is based on the residuals from the additive main effects submodel discussed earlier. In this case, the four multivariate analysis of variance statistics reported by the GLM procedure are based on different functions of the four positive characteristic roots of the HE^{-1} matrix.

Let $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 > 0$ denote the four positive characteristic roots of HE^{-1}. The Hotelling–Lawley Trace is $\Sigma_i\lambda_i$. Roy's Greatest Root is λ_1. Pillai's Trace is $\Sigma_i\lambda_i/(1 + \lambda_i)$. Wilks's Lambda is $1/\Pi_i(1 + \lambda_i)$. SAS Institute, Inc. (2004), reports that Roy's Greatest Root test "provides only a lower bound on the p-value." This can result in spuriously significant results. The other statistics are conservative in that they tend to overestimate the p–value: "the Hotelling-Lawley trace seems to be the least conservative." Volume 6 (p. 24) of Kotz and Johnson (1985) states that for large deviations from the null hypothesis, when the values of the characteristic roots of HE^{-1} are far apart, the Hotelling–Lawley trace statistic is more powerful than the other statistics.

When the sample sizes are large, the Wilks, Hotelling, and Pillai criteria are similar to one another. In this and in subsequent examples, the Hotelling–Lawley Trace statistic is used for illustrations. The Hotelling–Lawley Trace multivariate analysis of variance test of equality of treatment plant effects is performed using the multivariate analysis of variance features within the GLM procedure of the SAS System. In this example its distribution is approximated by an F distribution with 28 numerator and approximately 3 denominator degrees of freedom. The results of the SAS analysis are displayed in Table 3. The F value associated with the observed value of the Hotelling-Lawley trace is 81.45 and is significant at the .004 level.

The results of the 964 bootstrap resamples are summarized in Table 4. A portion of the listing of the empirical distribution function is shown in the table, and the closest location to the value associated with the observed F statistic, 81.45, is shown in bold. The observed bootstrap significance level is approximately 1 – .966 = .034. In agreement with the normal theory test, the treatment plants are significantly different. However the bootstrap significance level is an order of magnitude larger than the normal theory significance level, .034 vs. .004.

The differences among the treatment plants are clearly displayed in the principal components plots, Figure 4 in Feder et al. (2009). In that plot the treatment plants are separated from one another but the finished water samples and the distribution water samples from the same treatment plant are clustered together.

The empirical distribution function of the F statistic corresponding to the bootstrap resamples is displayed in Figure 2, along with the normal theory F distribution that is included for comparison. A reference line is shown at the F value associated with the observed value of the statistic (81.45), and the attained significance levels (%) are read from the vertical axis as 100 minus the values at which the reference line intersects the distribution functions. The bootstrap empirical cdf for the statistic lies below the normal theory cdf. This implies that the bootstrap empirical distribution of the F statistic has a heavier tail than the normal theory-based F distribution. This could potentially result in stating significance with the normal theory-based test but non-significance with the nonparametric bootstrap test. This was suggested in the present example, where the normal theory significance level (.004) indicates an order of magnitude more extreme level of significance compared to the bootstrap nonparametric significance level (.034).

FOUR-WAY ANALYSIS OF VARIANCE—U.S. ENVIRONMENTAL PROTECTION AGENCY AND ASSOCIATION OF METROPOLITAN WATER DISTRICTS 35-UTILITIES STUDY

Example—35-Utilities Study Data

The 35-utilities study data (U.S. EPA/AMWA, 1989) are more extensive than the Schenck et al. (2009) data (Bull et al., 2009b). The 35 utilities were sampled within relatively short time spans during four consecutive seasons. There were three different water sources: groundwater, lake reservoir, and flowing stream. Two disinfection processes were utilized: free chlorine and chloramination. All but three of the utilities had the

TABLE 3
Multivariate Analysis of Variance Test Statistic for Equality of Treatment Plants. Two-way matched pairs normal theory multivariate analysis of variance. Schenck and Sivaganesan (2009) data.

Statistic	Value	F-Value	Numerator degree of freedom	Denominator degree of freedom	Pr > F
Wilks' Lambda	0	18.56	28	12.239	<0.0001
Pillai's Trace	3.29	3.98	28	24	0.0005
Hotelling-Lawley Trace	506.82	81.45	28	2.5714	0.0042
Roy's Greatest Root	372.35	319.16	7	6	<0.0001

TABLE 4

Summary of F-Values Corresponding to $B = 964$ Bootstrap Resamples of the Multivariate Analysis of Variance Test Statistic for Equality of Treatment Plants. Two-way matched pairs nonparametric multivariate analysis of variance. Schenck and Sivaganesan (2009) data.

F-Value[a,b]	Cumulative frequency	Cumulative %
75.34	930	96.47
80.74	**931**	**96.58**
83.29	932	96.68
90.79	933	96.78

[a]Empirical distribution of ordered Hotelling-Lawley trace F-values based on bootstrap resampling iterations under the null hypothesis

[b]F-Value of Hotelling-Lawley trace based on original data (from Table 3) = 81.45

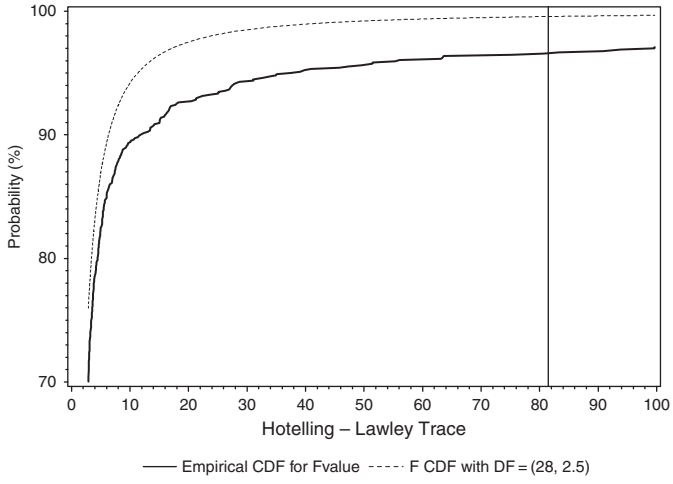

FIG. 2. Empirical CDF and normal theory F-distribution CDF for Hotelling-Lawley trace F-value for testing hypothesis of no plant treatment effects. Two-way matched pairs analysis. Significance levels (%) are 100 minus the values at which the vertical reference line crosses the curve. Schenck and Sivaganesan (2009) data.

same water source and used the same disinfection process for all four seasons. The other three differed in a single season. If, in one season, a utility altered its source or treatment process, it was treated for that season as if it was a different utility.

Based on the discussion in Bull et al. (2008a, 2008b), six DBP classes were measured for each season in each utility: trihalomethanes, haloacetic acid, haloacetonitriles, haloketones, trihaloacetaldehyde, and trichloropicrin. The data set also includes measurements of the characteristics of the intake water. Of particular interest are TOC and bromide level in the raw water.

The relationships between the disinfection by-products and potential health effects provide the basis for the selection of water characteristics for statistical comparisons to assess similarity among water supplies. The treated water chemical composition was summarized in an 11-dimensional vector. The components of the summary are molar concentrations (nmol/L) within the following 10 classes (as well as percent bromination within the total trihalomethane class):

1. Total trihalomethane.
2. Trihaloacetic acid.
3. Dihaloacetic acid.
4. Monohaloacetic acid.
5. Trihaloactonitriles.
6. Dihaloactonitriles.
7. Trihaloketones.
8. Dihaloketones.
9. Trihaloacetaldehyde.
10. Trihalopicrin.
11. Percent bromination within the total trihalomethane class (i.e., nmol/L bromine/total nmol/L).

Normal Theory Test

Statistical inferences were carried out in a manner directly analogous to those in the previous examples.

Let X_{hijk} denote the summary vector of water characteristics of the water sample, where the subscripts denote the following:

$$\begin{aligned} &h = \text{treatment} && h = 1,\ldots,H \\ & && (H = 2 \text{ in the } 35\text{-utilities study}) \\ &i = \text{source} && i = 1,\ldots,I \ (I = 3) \\ &j = \text{season} && j = 1,\ldots,J \ (J = 4) \\ &k = k(h,i) \text{ utility id} && k = 1,\ldots,K(h,i) \end{aligned}$$

It is assumed that

$$X_{hijk} \sim (\mu_{hijk}, \Sigma)$$

The model assumes a common 11×11 covariance Σ across the conditions. The chemical composition vector X_{hijk} is represented as

$$X_{hijk} = \mu_{hijk} + \varepsilon_{hijk}$$

where

$$\mu_{hijk} = \mu_0 + \alpha_h + \beta_i + \gamma_j + U_{k(h,i)}$$

subject to the conditions $\Sigma_h \alpha_h = \Sigma_i \beta_i = \Sigma_j \gamma_j = 0$; $\Sigma_{k(h,i)} U_{k(h,i)} = 0$ for all h,i. This is a four-way additive analysis of variance model. In this model the utilities are treated as fixed effects. That is, they are not considered to be a random sample from a

larger population of utilities and inferences are to be made about the specific 35 utilities included in the data.

The constraints imply that μ_0 represents the average response across treatments, water sources, seasons, and utilities; $\mu_0 + \alpha_h$ represents the average response for treatment type h averaged across water sources, seasons, and utilities; and $\mu_0 + \beta_i$ represents the average response for water source i averaged across treatments, seasons, and utilities; $\mu_0 + \gamma_j$ represents the average response for season j averaged across treatments, sources, and utilities; and $\mu_0 + U_{k(h,i)}$ represents the average response for utility $k(h,i)$ averaged across seasons.

An additive model is adopted for relative simplicity of illustration. Additivity implies that the effect of treatment is the same for all water sources and seasons, etc. Extensions of the model are considered in the discussion section later.

The following hypotheses can be tested based on the four-way multivariate analysis of variance model.

- No treatment effects:

$$H_0^{(\alpha)}: \alpha_h = 0 \text{ for all } h \text{ vs. } H_1^{(\alpha)}: \alpha_h \neq 0 \text{ for some } h$$

- No water source effects:

$$H_0^{(\beta)}: \beta_i = 0 \text{ for all } i \text{ vs. } H_1^{(\beta)}: \beta_i \neq 0 \text{ for some } i$$

- No season effects:

$$H_0^{(\gamma)}: \gamma_j = 0 \text{ for all } j \text{ vs. } H_1^{(\gamma)}: \gamma_j \neq 0 \text{ for some } j$$

- No utility (treatment, water source) effects:

$$H_0^{(U)}: U_{k(h,i)} = 0 \text{ for all } k(h,i) \text{ vs.}$$
$$H_1^{(U)}: U_{k(h,i)} \neq 0 \text{ for some } k(h,i)$$

The error sums of squares and cross-products matrix is based on the residuals from the additive model, which corresponds to the pooled two-, three-, and four-way interaction effects.

Nonparametric Bootstrap Test

The bootstrap analysis is based on the studentized residuals from the analysis of variance fit, in direct analogy with the procedure discussed for the matched-pairs analysis.

- Fit the full four-way additive multivariate analysis of variance model and calculate the studentized residuals from the fit:

$$\varepsilon_{hijk} = \left[diag(stderr(X_{hijk} - \hat{X}_{hijk})) \right]^{-1} (X_{hijk} - \hat{X}_{hijk})$$

where X_{hijk} denotes the model-predicted value and $\text{diag}(\text{stderr}(X_{hijk} - \hat{X}_{hijk}))$ is a diagonal matrix with diagonal elements the standard errors of the components of the residual vector.

- For the 35-utilities data, there is a potential maximum of $N = 35 \times 4 = 140$ studentized residuals. However, due to missing concentration values, only 108 studentized residual vectors were used in the analysis. Draw a random sample of size $N = 140$ with replacement from the studentized residuals. (There are 108^{140} such possible random samples.) Denote these residuals by ε_1^*, $\varepsilon_2^*, \ldots, \varepsilon_N^*$. These residuals satisfy the null hypothesis.
- Proceed in analogy to the matched-pairs analysis bootstrap test.

Example—35-Utility Study

The normal theory multivariate analysis of variance tests were carried out by performing multivariate analysis of variance using the GLM procedure of the SAS System (SAS Institute, Inc., 2004). Results are displayed for the following tests of hypothesis:

- Equality of water sources.
- Equality of seasons.
- Equality of treatments.
- Equality of utilities within treatments and sources (averaged over seasons).

These are summarized in Table 5.

The results of the bootstrap resamples of the F values associated with the Hotelling–Lawley trace statistic are summarized in Table 6. Portions of each of the listings of the empirical distribution functions are shown in the table. The observed normal theory F_0 values are included in the table for reference.

In each case, the normal theory tests indicate highly statistically significant differences. Each observed F statistic F_0 lies far out in the upper tail of the bootstrap empirical distribution. The normal theory test and the bootstrap test are in agreement.

DISCUSSION

This article extends the discussion in Feder et al. (2009) in two ways:

1. The statistical inference procedures in Feder et al. (2009) were based on normal theory multivariate analysis of variance. This article discusses nonparametric alternatives based on the "bootstrap" resampling technique.
2. The statistical analysis procedures in Feder et al. (2009) were illustrated with monitoring data from five water treatment plants. In addition to these data, the statistical analysis procedures in this article are illustrated with a data set based on waters from 35 utilities, which permits inclusion of a greater number of water constituents and greater structure in the models.

TABLE 5

Hotelling-Lawley Trace Multivariate Analysis of Variance Test Statistics for Equality of Sources, Equality of Seasons, Equality of Treatments, and Equality of Utilities. Four-way normal theory multivariate analysis of variance. 35-Utilities data (1989).

Test	Hotelling-Lawley Trace value	Associated F-value	Degrees of freedom of test	Significance level
H_0: No overall water source effect	2.88	7.36	(22, 93.7)	<0.0001
H_0: no overall season effect	1.57	2.67	(33, 124.8)	<0.0001
H_0: no overall treatment effect	1.18	6.11	(11, 57)	<0.0001
H_0: no overall utility effect	47.49	7.06	(374, 388.0)	<0.0001

TABLE 6

Summary of F-Values Corresponding to B = 1,000 Bootstrap Resamples of the Multivariate Analysis of Variance Test Statistic for Equality of Sources, Equality of Periods, Equality of Treatments, and Equality of Utilities. Four-way normal theory multivariate analysis of variance. 35-Utilities data (1989).

F-Value	Cumulative frequency	Cumulative %
H_0: no overall water source effect: $F_0 = 7.36$[a]		
3.95	997	99.7
3.99	998	99.8
4.12	999	99.9
4.38	1,000	100
H_0: no overall season effect: $F_0 = 2.67$[a]		
2.25	999	99.9
2.26	1,000	100
H_0: no overall treatment effect: $F_0 = 6.11$[a]		
5.13	998	99.8
5.19	999	99.9
H_0: no overall utility effect: $F_0 = 7.06$[a]		
2.09	998	99.8
2.13	999	99.9
2.21	1,000	100

[a]F_0 corresponds to the normal theory Hotelling-Lawley trace F-statistic value, shown in Table 5.

In this article, the bootstrap discussion is restricted to hypothesis testing applications in multivariate analysis of variance. The intent is to illustrate that the bootstrap resampling approach can be applied to comparisons of DBP among water supplies as alternatives to the standard multivariate normal theory procedures.

The bootstrap procedure can be extended to inferences beyond those discussed in this article. For example when comparing the chemical composition of finished water with that of distribution water in the Schenck et al. (2009) data, it was assumed that the covariance structure is the same for both of these groups. The bootstrap procedure can be used to develop robust tests for equality of variances (if sufficient data exist). Bootstrap procedures can also be applied to (1) inferences about principal components, (2) comparisons of average chemical composition of water supplies in the presence of heterogeneous covariance structure, and (3) robust outlier detection procedures.

The bootstrap procedure can be extended to construct univariate or multivariate confidence regions. This is a direct generalization of hypothesis testing in that the confidence region consists of the set of parameter values that would not be rejected by the associated hypothesis test. Thus, conducting a succession of bootstrap hypothesis tests with varying null hypotheses results in specifying a confidence region

This article develops and illustrates statistical methodology for assessing the similarity of alternative water supplies. Several of the statistical assumptions were made for the purpose of illustrating statistical procedures. If the methods are to be used for a rigorous risk assessment, sensitivity analyses should be conducted to assess the effects on inference results of variation in the assumptions. For example in the comparison of the treatment plant effects with the Schenck et al. (2009) data, the nonparametric bootstrap significance level was an order of magnitude larger than that of the parametric normal theory test. Thus in this inference results may be sensitive to the normality assumption. If the extent of the data permits, other model assumptions such as constant covariance or additive effects can be evaluated to determine their impact on analysis results. Provisions for the critical evaluation of modeling assumptions might be incorporated into the design of future treatment plant monitoring studies.

In the analysis of the 35-utilities study the utilities were treated as fixed effects, with inferences restricted to the specific 35 utilities that were included in the study. As an extension of the analyses considered in this article, additional analyses might be carried out treating the utilities as random effects, i.e., as a randomly selected sample from a larger

population of utilities. Inferences would apply to the larger population of utilities.

From an operational consideration in carrying out the analysis of variance tests, the principal difference between the random effects analysis tests and the fixed effects analysis tests discussed in this article is the choice of the appropriate sums of error sums of squares and cross-products matrix E. In a fixed effects analysis the error matrix is the residual matrix from the additive model. This aggregates all the two-factor and higher order interactions. In a random effects analysis the error matrix is based on interactions between the effect under consideration and the utilities in the sample. Interaction effects need to be added to the model. In the case of completely balanced data the appropriate error sums of squares and cross-product error matrix with which to test an effect is the two-factor interaction effect matrix between that effect and the utilities. In the case of unbalanced data the appropriate error sums of squares and cross-product error matrix is an appropriate linear combination of interaction matrices, depending on the design and the degree of imbalance. The construction of such tests in the multivariate situations is nonstandard and is not a readily available analysis in most statistical packages.

REFERENCES

Bull, R. J., Teuschler, L., and Rice, G. 2009a. Determinants of whether or not mixtures of disinfection by-products are similar. *J. Toxicol. Environ. Health A*. 72:437–460.

Bull, R. J., Rice, G., Teuschler, L., and Feder, P. 2009b. Chemical measures of similarity among disinfection by-product mixtures. *J. Toxicol. Environ. Health A*. 72:482–493.

Efron, B., and Tibshirani, R. J. 1993. *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Feder, P. I., Ma, Z., Bull, R. J., Teuschler, L. K., Schenck, K. M., Simmons, J. E. and Rice, G. E. 2009. Evaluating sufficient similarity for disinfection by-product (DBP) mixtures: Multivariate statistical procedures. *J. Toxicol. Environ. Health A*. 72:468–481.

Hall, P., and Wilson, S. R. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics* 47:757–762

Kotz, S., and Johnson, N. L. 1985. *Encyclopedia of statistical sciences*, vol. 6. New York: John Wiley and Sons.

Morrison, D. F. 1976. *Multivariate statistical methods*, 2nd ed. New York: McGraw-Hill. (4th ed., 2002, Brooks/Cole.)

Rice, G. E., Teuschler, L. K., Bull, R. J., Simmons, J. E., and Feder, P. I. 2009. Evaluating the similarity of complex drinking-water disinfection by-product mixtures: Overview of the issues. *J. Toxicol. Environ. Health A*. 72:429–436.

SAS Institute, Inc. 2004. *SAS/STAT® 9.1 user's guide*. Cary, NC: SAS Institute, Inc.

Schenck, K. M., Sivaganesan, M., and Rice, G. E. 2009. Correlations of water quality parameters with mutagenicity of chlorinated drinking-water samples. *J. Toxicol. Environ. Health A*. 72:461–467.

U.S. Environmental Protection Agency. 1986. *Guidelines for the health risk assessment of chemical mixtures*. EPA/630/R-98/002. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www.epa.gov/ncea/raf/pdfs/chem_mix/chemmix_1986.pdf.

U.S. Environmental Protection Agency. 2000. *Supplementary guidance for conducting health risk assessment of chemical mixtures*. EPA/630/R–00/002. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. http://www.epa.gov/ncea/raf/pdfs/chem_mix/chem_mix_08_2001.pdf.

U.S. Environmental Protection Agency/Association of Metropolitan Water Agencies. 1989. *Disinfection by-products in United States drinking waters*. Final report. La Verne, CA: James M. Montgomery Consulting Engineers and the Metropolitan Water District of Southern California.

Westfall, P. H., and Young, S. S. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustments*. New York: John Wiley and Sons.